Stealth Assessment of Problem-Solving Skills from Gameplay

Weinan Zhao, Valerie Shute, Lubin Wang
Florida State University
Tallahassee, FL
wz11@my.fsu.edu, vshute@fsu.edu, lw10e@my.fsu.edu

ABSTRACT

Stealth assessment represents a promising way to address the needs of validly measuring and supporting important 21st century competencies (e.g., creativity, problem solving) within interactive digital environments (e.g., video games). The assessment is woven into the environment such that it becomes invisible to students, which is conducive to eliciting targeted competencies (Shute, 2011). Stealth assessment also runs dynamically, enabling real-time support. We use ECD (Evidence-Centered Design, Almond, & Lukas, 2003) as our assessment design framework for creating stealth assessments that capture far more information related to multiple competencies compared to traditional forms of assessment, which typically report a single summative score, and/or or judgments of right or wrong. To date, we have developed a number of stealth assessments for use in different games to examine various competencies. For example, we have designed three stealth assessments to measure various cognitive and noncognitive variables in a game called Physics Playground (Shute & Ventura, 2013). The focal competencies included persistence (Ventura, Shute, & Small, 2014), qualitative physics knowledge (Shute, Ventura, & Kim, 2013), and creativity (Kim & Shute, in press). From these design and development efforts, we have learned a number of useful lessons about developing and applying stealth assessment. This will comprise the focus of our paper—lessons learned and best practices related to the design process of stealth assessment. We will demonstrate the process of designing stealth assessment using a research project that assesses problem solving skill in the popular game Plants vs. Zombies 2. Results from our evaluation study show that our game-based assessment is promising, correlating with the external measures of problem solving: Raven's progressive matrices (r = .40, p < .01) and MicroDYN (r = .48, p < .01). However a larger sample size is needed to establish definite claims about its validity.

ABOUT THE AUTHORS

Weinan Zhao is a PhD candidate of Instructional Systems and Learning Technology at the Florida State University. His research interests are game-based learning and assessment, knowledge modeling, educational data mining and intelligent tutoring systems.

Valerie Shute Valerie Shute is the Mack & Effie Campbell Tyner Endowed Professor in Education in the Department of Educational Psychology and Learning Systems at Florida State University. Her general research interests hover around the design, development, and evaluation of advanced systems to support learning--particularly related to 21st century competencies. Her current research involves using games with stealth assessment to support learning—of cognitive and noncognitive knowledge, skills, and dispositions.

Lubin Wang Lubin Wang is currently a doctoral candidate in Instructional Systems & Learning Technologies program at Florida State University. Her research interests include using the latest technologies to measure and support learning. She is particularly interested in the assessment and improvement of problem-solving skills and persistence, as well as identifying gaming-the-system behaviors during gameplay. She has participated in various research projects led by Dr. Shute, and is co-authoring several papers/book chapters with her.

Stealth Assessment of Problem-Solving Skills from Gameplay

Weinan Zhao, Valerie Shute, Lubin Wang
Florida State University
Tallahassee, FL
wz11@my.fsu.edu, vshute@fsu.edu, lw10e@my.fsu.edu

INTRODUCTION

Video games and simulations are considered to be good vehicles to improve students' 21st century competencies, such as creativity, problem solving, and persistence (Shute, Ventura, Kim, & Wang, 2014), which are regarded as important for students to be successful in 21st century (Partnership for 21st Century Learning, 2015). However, designers of such interactive digital environments are likely to face difficulties when they attempt to measure these competencies. For instance, 21st century competencies are often complex, thus very difficult to measure using traditional forms of assessment such as multiple-choice format. Moreover, assessment in gaming and simulation environments should provide real-time information for learning support purposes, which requires automated scoring. In addition, assessment in video games often needs to be carried out in an unobtrusive way so that the flow of gameplay will be maintained and the player's enjoyment of the game will not be affected. These factors call for a new way of obtaining information and making inferences about student's competency levels.

Stealth assessment (Shute, 2009) is an innovative way to address these issue. It uses ECD (Evidence-Centered Design; Mislevy, Almond, & Lukas, 2003) as the assessment design framework which: a) extends the traditional assessment space to embrace a broad range that includes complex competencies; b) provides a solid framework for the assessment designer to define the claims to be made, the evidence to be gathered, and the rules to score game actions and make inferences from the evidence through the claims; and c) puts no limitations on the format of the information to be gathered. Therefore with ECD as its basis, stealth assessment is designed to be invisible to students, run dynamically throughout gameplay, and enable real-time learning support. In addition, stealth assessments capture far more information related to multiple competencies (and their sub-skills) compared to traditional forms of assessment which are typically unidimensional and produce just a single summative score on a test or judgments of right or wrong.

We have developed a number of stealth assessments for use in different games to examine various competencies, such as persistence, qualitative physics knowledge, and creativity in game Physics Playground (Shute & Ventura, 2013; Ventura, Shute, & Small, 2014; Ventura, Shute, & Zhao, 2012; Shute, Ventura, & Kim, 2013; Kim & Shute, in press). In this paper, we summarize the lessons we have learned from our efforts designing and developing stealth assessment, in the form of a model of the design process. We will then demonstrate the process of designing stealth assessment using a current research project that assesses problem solving skills in the popular game Plants vs. Zombies 2.

EVIDENCE-CENTERED DESIGN AND STEATH ASSESSMENT

Evidence-centered design (ECD) is a framework that provides guidelines for assessment designers to create educational assessments in terms of evidentiary arguments (Mislevy, Almond, & Lukas, 2003). This helps to make valid inferences about the target competency level of a student based on evidence collected from his or her performance on particular tasks, according to the principles of evidentiary reasoning. Figure 1 shows the three main models (and the relationships among them) of ECD: the competency model, the evidence model, and the task model. The three interrelated models are the result of the design process.

The competency model includes a set of variables that flesh out the target competency. The competency variables are usually interconnected to reflect the structure of the competency. The task model specifies: a) the features of tasks that need to be created that will elicit students' performance related to the target competency, and b) the features of the work products as the outcome of the students' performance. The evidence model connects the competency and task models by introducing a) a set of variables called observables; b) a set of rules that specify how observables can be extracted from the work products specified in the task model; and c) a statistical model that connects the observables

with competency model variables. The ECD design process focuses on building these three models based on the purpose of the assessment. After the three models are specified, they are then implemented as an operational assessment. When delivered, the assessment machinery collects information of the work products when a student is performing the tasks. The collected information is used to extract observable values which in turn are imported to the statistical model, which calculates the estimated values of competency model variables. This process can be iterated when new work products are generated.

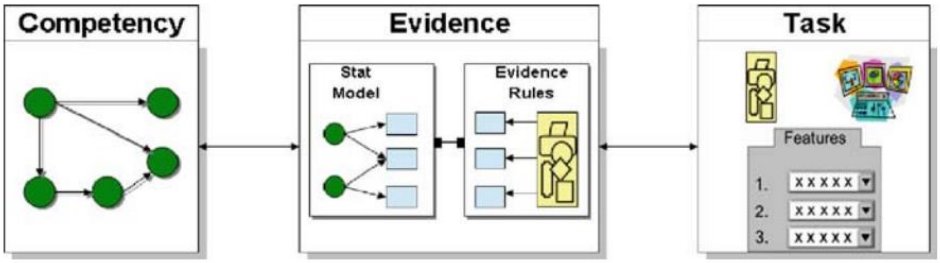

Figure 1. Three Main Models of ECD (adapted from Mislevy, Almond, & Lukas, 2003)

ECD supports a broad range of assessment types and addresses the recent needs in educational assessment to make inferences about a broad range of competencies, including higher-order competencies such as problem solving skills, creativity, critical thinking, and so on. Also ECD makes it possible to make valid inferences out of large amount of complex data from students, which can be captured when technologies (such as games and simulations) are employed in educational settings.

During the past decade, we have designed stealth assessments for different complex competencies within different games (e.g., Kim & Shute, in press; Shute & Kim, 2011; Shute, Ventura, & Ke, 2015; Shute, Ventura, & Kim, 2013; Ventura, Shute, & Small, 2014). We used video games as the vehicle for assessment because a) games are increasingly popular among teenagers (Lenhart et al., 2008); b) well-designed games engage students in enjoyable activities in meaningful contexts, enabling the assessment engine to capture cohesive, detailed information about players' target competency in an unobtrusive way (which can additionally serve to reduce anxiety often associated with traditional tests); and c) according to Gee (2003) and other researchers in this area, well-designed games can help students acquire important higher-order competencies. In short, stealth assessment is used to assess and update claims about a student's competencies continuously, capturing detailed information from gameplay and updating estimates of competency levels. The precision of the assessment is improved across time as more evidence is accumulated. The assessment results can also be used to provide learning support—either by the teacher or system.

LESSONS LEARNED: DESIGN PROCESS OF STEALTH ASSESSMENT

We encountered a number of issues while designing stealth assessments for various competencies in different games. After resolving these issues, we learned many useful lessons. We organize the issues/lessons learned with a process model shown in Figure 2. This is an adaption of the design process model of ECD (Mislevy, Steinberg, & Almond, 2003) and reflects the process we are currently employing in designing stealth assessment for video games.

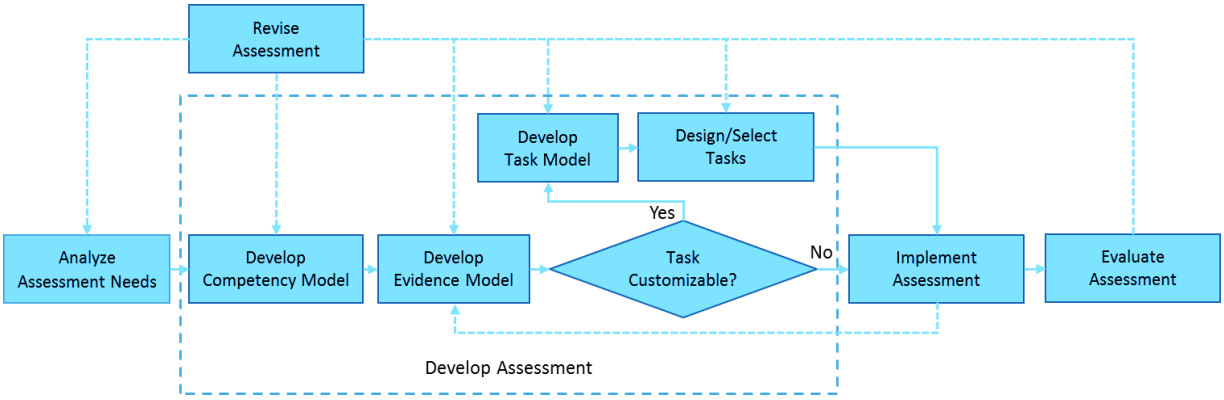

Figure 2. Design Process of Stealth Assessment

Analyze Assessment Needs

Like design processes in other areas, the first step in designing stealth assessment is to establish a clear understanding of the assessment needs, which will be used in each of the later stages of the design process. Questions need to be answered at this step include:

1) What is the target competency? Are there any facets of the target competency which are of particular interest?
2) What is the purpose of the assessment? Is it to be diagnostic/formative or summative?
3) Who is the target audience of the assessment?
4) Are the tasks in the game already in place? Are they customizable? Are tasks to be developed from scratch?
5) Is the telemetry (i.e., information captured in the log file) of the game customizable? If the telemetry is not customizable, what information is currently recorded?

A clear understanding of the answers to these questions at the beginning can help to improve the efficiency of the design process. For example, whether the telemetry of the game is customizable or not will influence later decisions when developing the evidence model. That is, if we developed the evidence model but discovered that there was no way to include relevant information required by the evidence model in the telemetry, then that would be a problem. We will describe in detail later how different pieces of information obtained in this stage can inform decisions made in the later stages.

Develop Competency Model

The purpose of this step is to identify the variables and their relationships in the competency model, based on the information collected about questions 1-3 in the first step. That is, by clearly delineating the target competency and target audience, this can help focus the work by the assessment designer. Also, having a clear understanding of the purpose of the assessment helps to set up the competency model to the right grain size (i.e., appropriate level). Having a larger (more general) grain size is usually sufficient for summative assessment purposes while specifying competencies to a smaller (more specific) grain size is often desirable for formative assessment purposes.

A valid competency model is often based on disparate sources, including theoretical arguments and empirical evidence from the literature, input from subject matter experts, and established standards, such as Common Core State Standards (CCSS Initiative, 2015) and Next Generation Science Standards (NGSS Lead States, 2013). When gathering information from these sources, useful information for later stages can also be collected, including:

1) Features of tasks that can inform the task model
2) Observables related to the target competency which are helpful in developing the evidence model
3) External measures of the same target competency which can be potentially used to evaluate the assessment

Keeping an eye on such information can save time/effort needed in later stages of the design process.

Develop Evidence Model

The central task of developing the evidence model is to identify the observables (in stealth assessment, we call them indicators) that can provide information for the assessment engine to estimate the values of the competency model variables. The information related to indicators (i.e., observables) collected in the previous step may serve as a reference in this task. Because the game title is usually decided at the beginning of the project (if the plan is to use an existing game as the assessment vehicle), the tasks within the game will inform or even decide the identification of indicators. Good indicators are usually derived from expert knowledge about the target competency, as well as an accurate, thorough task analysis of in-game activities. The latter often requires the assessment designer to spend a large amount of time playing the game and watching videos from expert players, if no game expert is present on the design team.

The range of information that can be captured from gameplay (usually in the form of telemetry) could put restrictions on defining the indicators. For example, consider one of our previous projects developing stealth assessments for three target competencies (i.e., problem-solving skill, spatial ability, and persistence) in a commercial game Portal 2 (Shute, Ventura, & Ke, 2015). We were not able to extract as many indicators as needed from the game because the log files were not customizable and their format made it very difficult to extract sufficient and meaningful information. The lesson learned here is to make sure that the log files are manageable or customizable as early as possible in your venture to design stealth assessment for a commercial game.

Once the indicators are identified, evidence rules need to be specified. As explained below, since we usually use Bayesian networks (Bayes nets) as the statistical model, we need to establish evidence rules that can extract values of the indicators in terms of discrete levels. Depending on the nature of each indicator, the number of levels for an indicator, as well as the rules to map the captured information to various states may vary. We will illustrate this with a concrete example in the next section. Note that a statistical model needs to be specified to connect the indicators with the competency model variables (e.g., if a player does X, this suggests that he is high on competency Y). We use Bayes nets as the statistical model because: a) it can calculate the estimated "value" of competency model variables in terms of the probability per level which fits with the needs of providing learning support; b) Bayes nets can model complex structures; c) they can incorporate expert knowledge (which often plays an important role when Bayes nets are initially constructed) as well as learn from data; and d) the computation of Bayes nets is fast and thus can work in real-time (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015). However, the work is often non-trivial since usually for each level of the game a different Bayes net needs to be constructed.

Develop Task Model

Tasks (i.e., levels or challenges) in a game may already be established or tasks may be customizable. For instance, many game titles ship with a level editor which enables the assessment designer to author customized levels, such as with Portal 2 and Physics Playground (previously Newton's Playground, Shute & Ventura, 2013). In such cases, we'll need to specify the task model in the form of the features of the tasks, such as target competencies or facets to address in the tasks, difficulty of the tasks, particular game mechanics, and so on.

Design/Select Tasks

Based on the task model, we are able to design and/or select the tasks we want to include in the operational assessment. A practical consideration is the length of the gameplay. It's important to note that employing stealth assessment in gameplay can result in an assessment that is longer than a traditional test, but the game is considerably more enjoyable than a typical test. Designers need to be careful to avoid player fatigue that may result from prolonged gameplay -- which could affect the results of the assessment. When we employ games as assessment vehicles, we typically limit gameplay to about 1-2 hours per session.

Implement Assessment

Implementation of stealth assessment involves modifying the code of the game to embed the process of capturing information needed to extract indicators, extracting indicators based on the evidence rules, and calculating the levels of the competency model variables based on the statistical model. A programmer who is familiar with Bayes nets will be a valuable asset to the design team.

Evaluate Assessment

The implemented stealth assessment needs to be evaluated to ensure that it meets the specific assessment needs, i.e., that it measures the target competency it is supposed to measure. External measures identified from the early stages of the design process need to be examined to decide whether they are appropriate for this evaluation step. Due to the nature of the stealth assessment, it is fairly easy to find an external measure which claims to measure the same target competency but in fact may not be well aligned with the target competency. One of the external measures we used in the Portal 2 project suffered from this issue. Therefore designers of stealth assessment need to make sure that the selected external measures align with the stealth assessment under evaluation. Once an aligned external measure (or measures) are identified, they need to be administered to the students who are playing the game with the stealth assessment. Correlations can be computed to see the degree to which the in-game (stealth) measures correlate with the external measure(s).

Revise Assessment

The results of the evaluation may suggest necessary revisions to the stealth assessment models. Based on the data collected in the evaluation study, revision may involve one or more earlier processes and its products. For instance, tasks may need to be redesigned to better elicit the indicators. Evidence rules for indicators may need to be modified to better reflect the differences among different levels of performance, and the structures and/or parameters of the Bayes nets may be adjusted.

Like design processes in other areas, the design process of stealth assessment is an iterative process and the quality of the assessment is improved through iteration. The iteration may happen among earlier stages. For example, some indicators built into the evidence model(s) may be impractical in the implementation stage due to computational complexity. If so, then the designer will need to go back to revise the evidence model(s).

Next we describe the design process of stealth assessment in the context of a research project that assesses problem solving skill in the popular game Plants vs. Zombies 2 (PvZ2).

ASSESS PROBLEM-SOLVING SKILLS IN PLANTS VS. ZOMBIES

PvZ2 is a tower defense game published by Electronic Arts. Players plant different kinds of plants in a limited space to defeat invading zombies. The game provides a rich environment to elicit players' performance in solving the problems – to defeat zombies with different resources, limitations and additional goals in different levels. Figure 3 shows a screenshot of PvZ2.

Figure 3. Screenshot of PvZ2

During initial communications with our collaborators at the GlassLab (http://about.glasslabgames.org/), we captured the following information regarding the needs of the assessment:

1) Target competency is problem-solving skill
2) Target audience includes middle school students
3) Glasslab has obtained the source code of the game, so we were able to select particular levels to be used for assessment purposes
4) The telemetry can be programmed/customized to capture any events happening in the game

This preliminary information was crucial for the later stages of the design process. However, information about the *purpose* of the assessment was not captured at that time. This resulted in our having to go back to revise the competency model after we finished identifying relevant indicators. That is, the GlassLab wanted to embed stealth assessment in the game to provide middle school teachers with real-time reports about students' problem-solving skills that are *aligned with the CCSS for Mathematical Practice* (Common Core State Standards Initiative, 2015). Because we weren't initially aware of this need for CCSS alignment, we built a competency model based on complex problem solving research. But when the actual purpose was made explicit, we revised the competency model. After reviewing the CCSS for Mathematical Practice, in conjunction with the problem solving literature, we derived four main facets of problem solving:

1) Understanding the givens and constraints in a problem
2) Planning a solution pathway
3) Using tools effectively/efficiently during solution attempts
4) Monitoring and evaluating progress

Next we developed the evidence model. The descriptions per standard in CCSS included example tasks. Students' success on those tasks indicates that they reached the level of proficiency the standard requires. The example tasks served as a good reference for our indicator identification process. That is, the tasks in the game have already been created, so we decided to focus on the first two worlds of the game comprising about 48 tasks in total. We identified relevant indicators through looking at similarities between the example tasks and the tasks in the game. To do this, we spent time playing through the game, as well as watching videos on YouTube for solutions of some particularly difficult levels. As a result, we identified a set of indicators for each facet of the problem-solving skills. Figure 4 illustrates some of the indicators associated with the competency model variables.

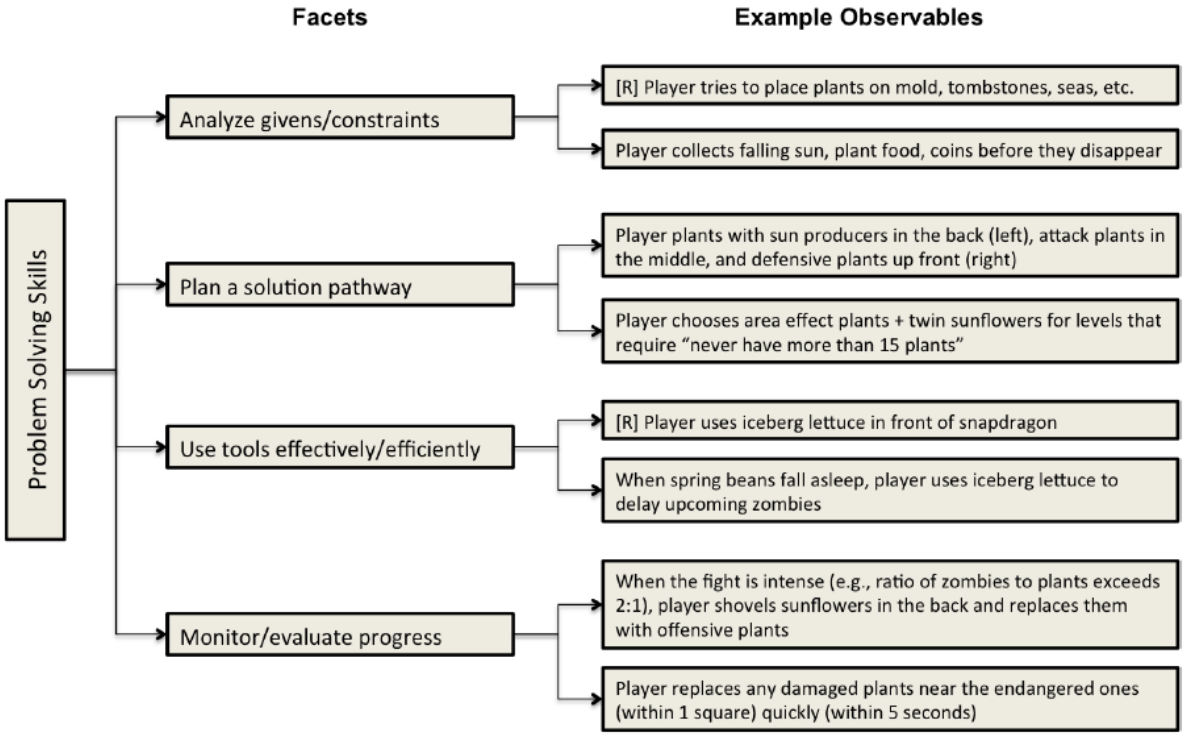

Figure 4. Competency Model of Problem-Solving Skills and Some of Related Indicators

We developed scoring rules based on the different characteristics of the indicators. For example, most indicators can be captured a number of times during a game level, and the time spent on good actions divided by total actions indicate the player's level of proficiency on a particular facet. For example, consider the indicator "*player collects falling sun*" (which is a good action in the game). Dividing the player's actual number of suns collected by the total number of suns that fell in a level suggests the degree of proficiency on that indicator. The degree to which the player does well (or poorly) informs the facet "analyze givens/constraints". Therefore, for each indicator of this kind, we calculated the ratio of good performance to total, and then converted the proportion of it into four discrete levels using cut scores. For a few indicators, it only matters if the player performed the corresponding action or not, such as "*Player uses iceberg lettuce in front of snapdragon*". In this case we simply scored it as yes or no.

To connect the indicators with the facets, we constructed a Bayes net for each level of the game. We did this because different levels have different indicators. Figure 5 illustrates the Bayes net we created for Day 1 of the first world (i.e., *Ancient Egypt*) using the Bayes Net software tool Netica (by Norsys Software Corporation). There are two main steps to build a Bayes net for a level:

1) *Identify the structure*. In this step we examined each indicator of the level to find out which facet(s) are linked to it (i.e., which facet(s) the indicator can provide evidence to make inferences related to the facet). We a) went through the full list of the indicators we came up with to identify those can be captured in the level; and then b) determined whether the identified indicators were connected or not to each of the facets. Towards that end, we created a Q-matrix (Suppes, 1969) to display the results of our analysis. The Q-matrix can then be converted to a Bayes net in Netica.
2) *Estimate the parameters*. To reason in a Bayes net from an indicator node to its parent facet(s), or from a facet node to its parent node (i.e., the node representing problem-solving skills), we needed to construct conditional probability tables (CPTs) per indicator node and facet node. A CPT specifies the probability of a node being at each level given its parent(s) being at various levels (e.g., low, medium, and high). To simplify the work we used a discrete graded response model (Muraki, 1992) which takes a set of IRT-like (IRT stands for item response theory, Hambleton, Swaminathan, & Rogers, 1991; Thissen & Wainer, 2001) parameters to generate the CPT. Thus, for each parent of the node, we estimated: a) a discrimination value, which

represents the discriminative ability of an indicator regarding the parent; and b) a set of difficulty values. Each value represents the particular difficulty of the indicator relative to a level. Finally, we estimated the marginal probability for the root node "ProblemSolvingSkills." This work was done through the collaboration among a learning scientist, a game expert, and a psychometrician.

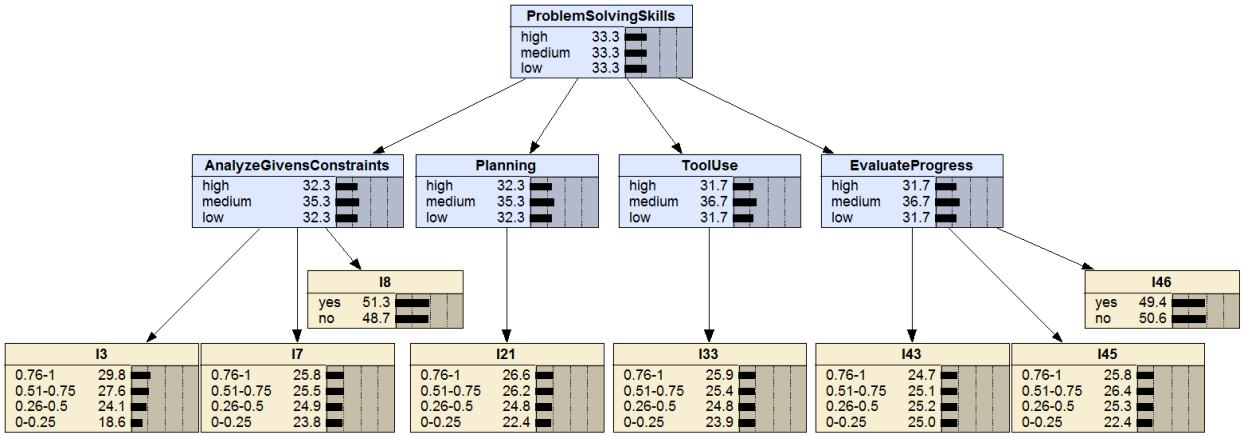

Figure 5. Bayes Net for Day 1 of Ancient Egypt—Prior Probabilities

Because we could not modify the existing levels or create new levels, we skipped the process of developing the task model. However, we removed a couple of the existing levels from gameplay as they did not provide any information for our assessment purpose. Next, the team at GlassLab modified the telemetry to capture the information required to extract and log all of the indicators. During this process, we revised the scoring rules for some of the indicators to make them implementable.

As soon as the build of PvZ2 with modified telemetry was ready, we conducted a small pilot study to a) evaluate our game-based assessment, and b) refine it with actual data. We selected Raven's Standard Progressive Matrices (Raven, 1941) and MicroDYN (Wustenberg, Greiff, & Funke, 2012) as our external measures to establish construct validity. Raven's Progressive Matrices measures eductive and reproductive ability, i.e., the ability to make meaning out of confusion and to generate high-level schema to handle complexity, by requiring participants to infer the pattern of the missing piece from the given patterns (Raven, 2000). MicroDYN is a performance-based assessment that requires the participants to figure out the cause relationships in real-world systems and to control such systems by manipulating the variables based on given causal models. Based on a task analysis of the game, we expected that both measures could capture some aspects of problem-solving skills that are needed to succeed in PvZ2, while MicroDYN would show a better alignment with our in-game measure since it focuses on facets related to dynamic problem-solving skills.

In November 2014, we tested ten undergraduate students who played PvZ2 for 90 minutes and then completed MicroDYN (30 minutes) and Raven's (15 minutes). We selected undergraduate students as our participants as a convenience sample and believed that undergraduate students would resemble middle school students on the target competency and facets. To compare the results of our game-based assessment with that of MicroDYN, we calculated Expected A Posteriori (EAP) value of the problem-solving skills node for each student with its probabilities at the end of gameplay for the three states (i.e., high, medium, and low), according to the formula: $EAP = 1 * P(High) + 0 * P(Med) + (-1) * P(Low)$.

Initially, we did not find significant correlations between our game-based assessment and the external measures. Therefore, we re-examined each part of the initial assessment model based on the in-game data we obtained. There were no significant problems in the scoring rules or the structures of the Bayes nets, so we focused on adjusting the parameters of the Bayes nets to improve the assessment model. For example, if the mean score of an indicator was low, we would increase the difficulty values associated with its CPT. After revising the Bayes nets, results showed that our game-based assessment was significantly correlated with MicroDYN (r = .74, p = .03), but not correlated with Raven's. This was in line with our expectation (i.e., Both measures could capture some aspects of problem-solving skills that are needed to succeed in PvZ2, while MicroDYN would show a better alignment with our in-game measure since it focuses on facets related to dynamic problem-solving skills). However, due to the very small sample

size and potential differences between the participants and the target audience, we still needed to verify the stealth assessment with a larger sample of the target audience before making any claims about the validity of the stealth assessment. Consequently, we conducted an evaluation study in May of 2015. Fifty-two students were recruited from a middle school in a Midwestern state of U.S. They played PvZ2 for three hours across three days. On the fourth day, they completed the MicroDYN test, followed by a demographics questionnaire and Raven's Progressive Matrices. After removing names of students whose data were missing and/or incomplete, we ended with a set of data from 47 students. Results showed that our game-based assessment of problem solving skills was significantly correlated with Raven's ($r = .40$, $p < .01$) and MicroDYN ($r = .48$, $p < .01$) scores.

SUMMARY AND DISCUSSION

In this paper we summarized the design process of stealth assessment, the lessons learned, and some best practices related to this process based on our experiences creating stealth assessment for different 21st century competencies in different video games. We demonstrated the process with a real example of designing stealth assessment for problem-solving skills in PvZ2. As indicated in the demonstration, although Bayes nets can be crafted based on experts' input, actual data plays an important role in refining the Bayes nets. There are a couple of algorithms which automate the process of revising Bayes nets according to actual data (i.e., machine learning). We did not cover these algorithms in the current paper, but those who are interested can refer to Almond et al. (2015).

In terms of limitations to the research presented here, our initial plan for the validation study was to recruit at least 100 students. Due to logistical issue, we tested a much smaller sample which does not allow us to make definite claims about the validity of our game-based assessment, despite the significant correlations between our stealth assessment estimates of problem solving skill and both of the external measures. However, we do plan to test the game with a larger sample of middle school students in the near future.

In conclusion, there are a number of features of stealth assessment that are quite promising, such as the ability to capture complex constructs, and run invisibly and dynamically while a student/player is immersed in gameplay. This suggests that it is a promising way to measure higher-order competencies and serve as the basis to enable learning support given current estimations of competency states. We hope that our experiences shared here can provide some guidance to those who are interested in designing stealth assessment in their games or simulations.

ACKNOWLEDGEMENTS

We are in debt to Russell Almond who helped us resolve both theoretical and practical issues in building the Bayes nets. We would like to thank the GlassLab team who are supporting our work assessing problem solving in Plants vs. Zombies 2—specifically Jessica Lindl, Liz Kline, Michelle Riconscente, Ben Dapkiewicz, and Michael John.

REFERENCES

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment.* Springer.

Common Core State Standards Initiative. 2015. http://www.corestandards.org/.

Common Core State Standards Initiative. (2015). Standards for mathematical practice. Retrieved from http://www.corestandards.org/Math/Practice/.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Kim, Y. J. & Shute, V. J. (in press). Opportunities and challenges in assessing and supporting creativity in video games. To appear in J. Kaufmann & G. Green (Eds.), *Research frontiers in creativity*. San Diego, CA: Academic Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Lenhart, A., Kahne, J., Middaugh, E., Macgill, A., Evans, C., & Vitak, J. (2008). Teens, video games, and civics. *Pew Internet & American Life Project, 16.*

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement, 16*(2), 159{176. doi: 10.1177/014662169201600206

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report RR-03-16). Princeton, NJ: Educational Testing Service. Retreived from http://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisci-plinary Research and Perspective, 1*(1), 3-62.

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* Washington, DC: National Academies Press.

Partnership for 21st Century Learning. 2015. http://www.p21.org.

Raven, J. C. (1941). Standardization of progressive matrices, 1938. *British Journal of Medical Psychology, 19*(1), 137-150.

Raven, J. (2000) The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology, 41,* 1–48.

Shute, V. J., & Kim, Y. J. (2011). Does playing the World of Goo facilitate learning? In D. Y. Dai (Ed.), *Design research on learning and thinking in educational setting: Enhancing intellectual growth and functioning* (pp. 359-387). New York, NY: Routledge Books.

Shute, V. J. (2009). Simply assessment. *International Journal of Learning, and Media, 1*(2), 1-11. doi: 10.1162/ijlm.2009.0014

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V. J., Masduki, I., Donmez, O., Kim, Y. J., Dennen, V. P., Jeong, A. C., & Wang, C-Y. (2010). Modeling, Assessing, and Supporting Key Competencies Within Game Environments. In D. Ifenthaler, P. Pirnay-Dummer, N. M. Seel (Eds.), *Computer-based diagnostics and systematic analysis of knowledge* (pp. 281-309). New York, NY: Springer-Verlag.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education, 80,* 58-67.

Shute, V. J. & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment.* Cambridge, MA: The MIT Press.

Shute, V. J., Ventura, M., & Kim, Y. J. (2013). Assessment and learning of qualitative physics in Newton's Playground. *The Journal of Educational Research, 106,* 423-430.

Shute, V. J., Ventura, M., Kim, Y. J. & Wang, L. (2014). Video games and learning. In W. G. Tierney, Z. Corwin, T. Fullerton, and G. Ragusa (Eds.), *Postsecondary play: The role of games and social media in higher education.* (pp. 217-235). Baltimore, MD: John Hopkins University Press.

Suppes, P. (1969). Stimulus response theory of finite automata. *Journal of Mathematical Psychology, 6,* 327-355.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ventura, M., Shute, V. J., & Small, M. (2014). Assessing persistence in educational games. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), *Design recommendations for adaptive intelligent tutoring systems: Learner modeling, Volume 2.* (pp. 93-101). Orlando, FL: U.S. Army Research Laboratory.

Ventura, M., Shute, V. J., & Zhao, W. (2012). The relationship between video game use and a performance-based measure of persistence. *Computers and Education, 60,* 52-58.

Wustenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—more than reasoning? *Intelligence, 40,* 1-14.